eGovernement-Thesaurus und Suchapplikationen in Vorarlberg für Landes- und Gemeindeparlamente und Landeshomepage

Dipl.-Inf.wiss. Manfred Hauer M.A.
Head
AGI – Information Management Consultants
Mandelring 238 b
67433 Neustadt an der Weinstrasse
Manfred.Hauer@agi-imc.de

Abstract: Semantische Netze und Bäume. computerlinguistische und statistische Textanalayse und Information Retrieval auf aussagekräftigen Medatdaten und Volltexten helfen Bibliotheken, Archive und Websites besser nutzbar zu machen. An drei Applikationen hat das Land Vorarlberg in Österreich Erfahrungen gesammelt.

1 eGovernment braucht Information Retrieval

eGovernment hat gewiß viele Aspekte, einer davon ist die Publikation von Dokumenten und eine möglichst effiziente Navigation und professionelle Suche in diesen Dokumentenbeständen. Die Grundidee vieler CMS-Projekte ist, mit drei Mausklicks zum Ziel zu kommen. Dies galt einst auch für die Landeshomepage des kleinen Bundeslandes Vorarlberg. Die Anzahl der HTML-Seiten und darin verlinkten PDF- oder MS-Office-Dokumente wuchs flott und so mancher Nutzer braucht weit mehr als drei Navigationsschritte. Das ist eigentlich logisch, denn: in Navigationsbäumen gibt es die goldene Regel, nicht mehr als 7 Auswahlen gleichzeitig anzubieten. Mit 7 x 7 x 7 = 343 Webseiten ist diese Vision gut einlösbar. Doch welche wichtige Kollektion ist so klein? Schon zu beginn der CMS-Implementierung wurde eine Volltextsuche hinterlegt, welche mit wachsenden Mengen immer weniger zufriedenstellene Resultate liefern konnte. Die Landesinformatik wurde für das Thema Retrieval sensibel und fand Ansatzpunkte in der Vorarlberger Landesbibliothek (VLB).

2 Mehr Bildung durch bessere Wissensspeicher

Die Vorarlberger Landesbibliothek begann 2002 ein Projekt zur besseren Inhaltserschließung und Suche, indem Inhaltsverzeichnisse digitalisiert und mit computerlinguistischen Methoden ausgewertet werden. Links auf die PDFs der Inhaltsverzeichnisse und die linguistisch extrahierten Deskriptoren (terminologisch kontrollierte und freie Sachbegriffe, Phrasen, Personen, Orte) wurden in das Bibliothekssystem zusätzlich eingespielt. Die linguistische Analyse hatte alle Deskriptoren auf ihre grammatikalische Grundform reduziert – das was Benutzer auch normalerweise eingeben – und durch Gewichtsalgorithmen bewertet. Die Gewichtung war im Ansatz nicht mit dem Bibliothekssystem umsetzbar, zumal Bibliothekssysteme bisher in der Regel überhaupt kein Relevance Ranking unterstützen, sehr wohl aber schnelle Sortierverfahren (z.B. nach Publikationsjahr). Die Benutzer konnten jetzt endlich Spezialthemen finden, die nur in den Inhaltsverzeichnissen erwähnt wurden, aber der von Google gewohnte Komfort des Relevance Rankings fehlte.

2003 startete eine Suchmaschine unter dem Namen „dandelon.com“, welche die wichtigsten Metadaten des Bibliothekssystems, aus der Digitalisierung und aus der maschinellen Indexierung als auch die suchbaren Dateien vereinte. Zusätzlich enthielt „dandelon.com“ schon am Start die Funktion, ein oder viele Thesauri gleichzeitig in die Suche mit einzubinden. Synonyme, Übersetzungen und Unterbegriffe erwiesen sich als die brauchbaren zusätzlichen Begriffe. Die mittlerweile 1.7 Millionen vielfältig vernetzten Begriffe aus 25 Sprachen werden auf die über 1 Million Buch- und Aufsatzdatensätze bei der Recherche angewendet. Somit ist die Suche semantisch unterstützt und sprachübergreifend, crosslingual, zumindest wenn eine Übersetzung bekannt ist. Alle Thesauri stammen aus speziellen Fachgebieten und sind vom Benutzer aus- bzw. einschaltbar. Der Erfolg dieser Suchmaschine und die Vernetzung mit dem Bibliothekssystem dieser und zahlreicher anderer Bibliotheken ließ auch die Landesinformatik aufhorchen.

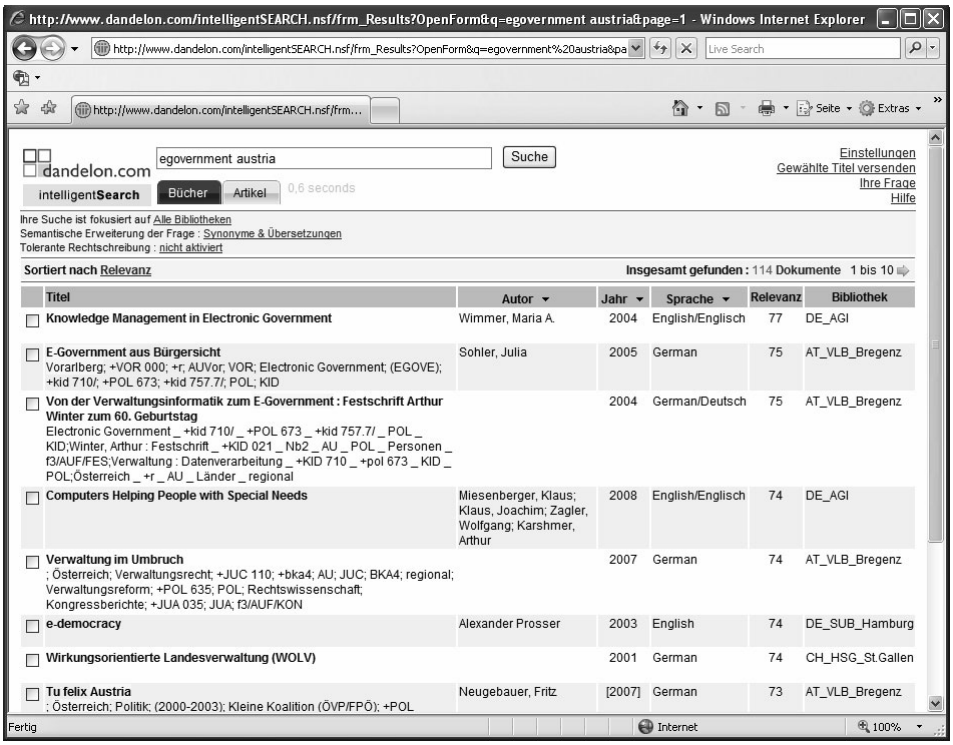

Abbildung 1: Schreibvarianten und Übersetzungen führen zu Resultaten, die ohne Semantik und Linguistik nicht möglich wären.

3 Landtagsinformationssystem (LIS) Vorarlberg

Vorarlberg legt großen Wert darauf, im eGovernment-Ranking innerhalb von Österreich positiv aufzufallen. (Österreich insgesamt liegt im europäischen Ranking vorne). Innerhalb des CMS mit den Inhalten der Landeshomepage gab es seit Jahren eine Suchfunktion in Word-Dateien, welche Sitzungsprotokolle enthalten. Regierung, Verwaltung, Parteien, Interessenvertretungen und Bürger waren damit nicht ganz glücklich, weshalb die in der Landesbibliothek bewährte Technik auch hier zum Einsatz kommen sollte – und seit Anfang 2007 produktiv ist. Die Protokolle sind stets aktuell und derzeit rückwirkend bis 1979 komplett online. Noch ältere Landtagsperioden sind in Arbeit und folgen und müssen wie bei den Büchern in dandelon.com über Digitalisierung und OCR verarbeitbar gemacht werden.

Das neue LIS (Landtagsinformationssystem) bietet den Einstieg über Top-Down-Navigation chronologisch und thematisch. Diese Verzeichnisse dienen dem Browsing im Datenbestand, ohne gezielt Fragen zu stellen. Eine Vernetzung zwischen Dokumenten, eine durchaus sinnvolle Erweiterung, gibt es nicht, denn eine solche könnte nur mit viel Aufwand und viel Kompetenz händisch gepflegt werden – und kann zukünftig wichtige Querbeziehungen kaum antizipieren.

Wichtig vor der Einführung war, dass der Personalaufwand so gering wie möglich sein sollte. Alle Themenpunkte einer parlamentarischen Sitzung werden einzeln als Word-Datei im Sekretariat des Landtagspräsidenten protokolliert, wie schon in allen Jahren zuvor. Das neue LIS beobachtet Dateiverzeichnisse und konvertiert vollautomatisch die Word-Dateien ins PDF-Format und übergibt die darin enthaltenen Texte der maschinellen Indexierung. Das Sekretariat erfasst wie bisher – nur in einer neuen mit Auswahlfeldern unterstützten Masken die Standard-Metadaten wie Parlamentsperiode, Sitzung, Tagesordnungspunkt, Datum der Sitzung, Typ des Dokuments. Parallel dazu vergibt ein Politikwissenschaftler der Landesbibliothek Einträge im Feld „Wichtige Personen“ und „wichtige Parteien“ – in der Regel kommen fast immer alle Parteien bei einem Tagesordnungspunkt zur Sprache, nur „wichtig“ hat hier nichts mit Häufigkeit oder Vorkommen an sich zu tun und kann mit maschinellen Verfahren nicht bewältigt werden. Dazu braucht es eine umfassende Weltsicht und gutes politikwissenschaftliches Verständnis.

Jede Anfrage im typischen „Suchschlitz“ wird möglichst zusätzlich in die EU-Sprachen übersetzt und mit Synonymen ergänzt und diese Query-Expansion über den Suchergebnissen angezeigt. Die Anzahl der zuschaltbaren Thesauri ist im Prinzip nicht begrenzt, konkret sind derzeit vier Thesauri aktiviert. Einerseits der Thesaurus der Europäischen Union, er ist in Brüssel und in vielen nationalen Parlamenten im Einsatz; dann der neu entwickelte eGovernment-Thesaurus und die Bibliothekssystematik der VLB und der schweizer Rechtsthesaurus JuriVoc. Zwei weitere Standardthesauri aus Sozial- und Wirtschaftswissenschaft folgen demnächst. Steht z.B. in einem Protokoll das Wort „Landebahnen“, so ergänzt die maschinelle Indexierung die Grundform „Landebahn“ und weiss entweder über ihren eigenen Thesaurus, dass solche im Kontext von „Flughafen“ vorkommen. Oder die externen, vom Benutzer steuerbaren Thesauri kennen die Verbindung. Tatsächlich führt die Eingabe „Airport“ in Bregenz zu Protokollen über Beratungen zum Ausbau von Landebahnen.

Wird so eine Suche noch kombiniert mit den terminologisch kontrollierten Feldern wie Partei oder Landtagsperiode oder Dokumententyp lassen sich sehr schnelle alle Anfragen einer Partei zum Thema „Airport“ in einer Periode ermitteln und nach Relevanz sortiert zeigen (neben anderen Sortieroptionen).

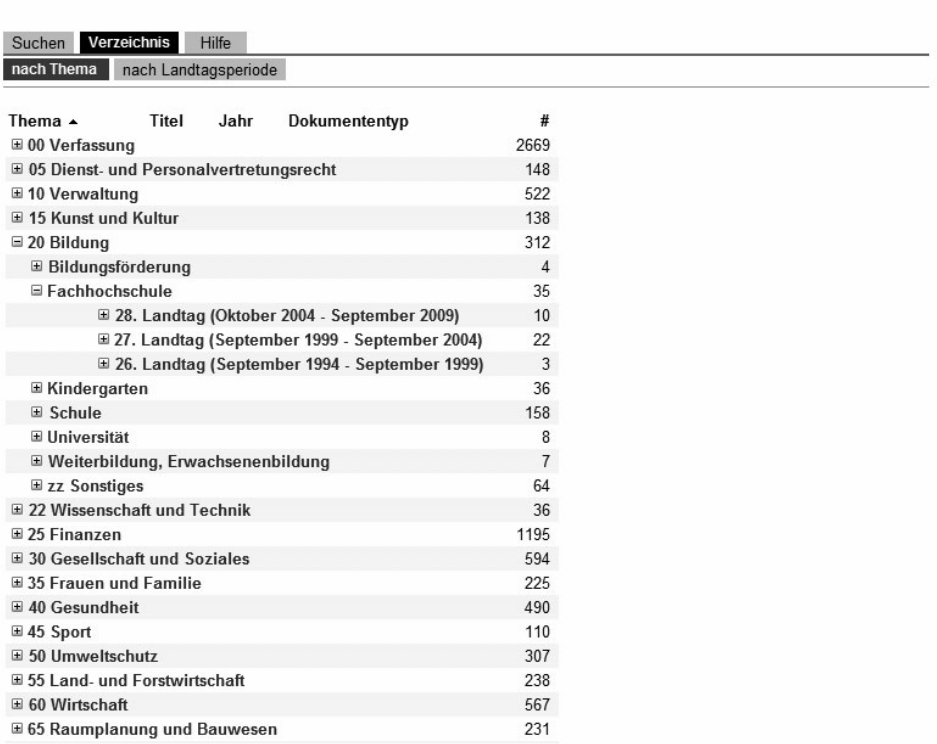

Abbildung 2: Top-Down-Navigation im Themenbaum

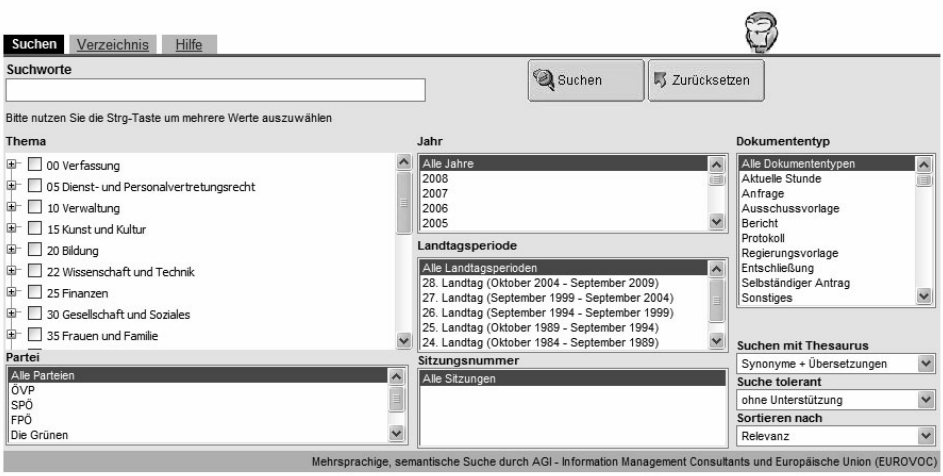

Abbildung 3: Suchmaske des Landtagsinformationssystems Vorarlberg (Ausschnitt)

Landtag · Recherche

Vorarlberger Landtag

Recherche in den parlamentarischen Materialien

Ihr Suchergebnis

Query : (Hochwasser OR ("Überschwemmung" OR "Hochwasser" OR "Überflutung" OR "Hochwasserschutz")) AND FIELD fd_PeriodeOfParliamentTX Contains "28. Landtag (Oktober 2004 - September 2009)" AND FIELD fd_TypeOfDocumentTX Contains "Anfrage"

zurück **Insgesamt gefunden:** 19 Dokumente 1 bis 10

Titel	Typ	Landtagsperiode	Sitzung	JahrAnfrageNr.
Das Hochwasser und die Folgen	Anfrage	28. Landtag (Oktober 2004 - September 2009)	2006-04	2006 29.01.122
Integraler Hochwasserschutz in Vorarlberg: Voraussetzungen für die Umsetzung schaffen!	Anfrage	28. Landtag (Oktober 2004 - September 2009)	2007-03	2007 29.01.206
Hochwasserschutz und finanzielle Absicherung bei Elementarereignissen	Anfrage	28. Landtag (Oktober 2004 - September 2009)	2006-03	2006 29.01.116
Parteipolitik auf dem Rücken von Hauswasseropfern?	Anfrage	28. Landtag (Oktober 2004 - September 2009)	2006-02	2006 29.01.098
Konsequenzen aus der Hochwasserkatastrophe vom 22./23. August 2005 in Vorarlberg	Anfrage	28. Landtag (Oktober 2004 - September 2009)	2005-07	2005 29.01.065
Welchen Gemeinden geht es schlecht in Vorarlberg?	Anfrage	28. Landtag (Oktober 2004 - September 2009)	2007-08	2007 29.01.230
Exkursion zum Lötschberg-Basistunnel - Erkenntnisgewinn oder Betriebsausflug	Anfrage	28. Landtag (Oktober 2004 - September 2009)	2005-08	2005 29.01.070
Exkursion zum Lötschberg-Basistunnel - Erkenntnisgewinn oder Betriebsausflug	Anfrage	28. Landtag (Oktober 2004 - September 2009)	2005-08	2005 29.01.070
Welchen Gemeinden geht es schlecht in Vorarlberg - Teil II	Anfrage	28. Landtag (Oktober 2004 - September 2009)	2007-08	2007 29.01.232
Verpachtung (Vermietung) von öffentlichem Wassergut	Anfrage	28. Landtag (Oktober 2004 - September 2009)		2008 29.01.303

Abbildung 4: Die Query-Expansion ist nachvollziehbar (und in der Suchmaske steuerbar)

LIS fand bei alle Nutzergruppen einen guten Anklang und strahlte mittlerweile auf eine zweite, diesmal privatwirtschaftlich getragene Lösung für die Gemeinden zunächst in Vorarlberg aus: www.gemeindedokumentation.at. Die Protokolle selbst von einem 300 Seelen-Dorf unterscheiden sich formal nicht von Land oder Bund.

Eine Projektgruppe der FH Burgenland in Eisenstadt verglich im Sommersemester 2008 insgesamt 17 Parlamentssysteme von Österreichs Bundesländern, einigen große Städten und dem umliegendenden, nicht deutschsprachigen Ausland. Gemessen wurde die Retrieval-Leistung und Funktionalität und die Zufriedenheit von Anwendern. Dabei landete das LIS Vorarlberg vor der Stadt Wien – und diese beiden deutlich vor den 15 anderen Angeboten.

4 Suche in der Landeshomepage www.vorarlberg.at

Weil LIS klaglos funktioniert, beauftragte 2007 die Landesinformatik erneut das Unternehmen des Autors, für deren Web-Angebot eine neue Suche zu entwicklen. Die erste Ausbaustufe ist online und eine zweite in Arbeit. Neu und anders als bei LIS, musste die Entwicklung sich streng an die Barrierefreiheitsnorm und die Gestaltungsvorgaben des Landes halten. Die aktuelle Suche ist äußerlich unscheinbar – ein kleiner Standards-Suchschlitz oben rechts auf jeder Seite und dann ein Suchergebnis, indem von Dokumentenoptimierung, thesaurusbasierter Suche, von maschineller Indexierung, von Optimierung des Relevance Rankings nichts direkt zu sehen ist. Die Antwort kommt schnell und gut nach Relevanz sortiert.

Abbildung 5: Suchbeispiel aus der Landeshomepage

Technisch werden die im CMS gepflegten Seiten im Betriebssystem eines Linux-Servers zur Anzeige abgelegt. Von dort holt sich die auf Lotus-Domino basierende Anwendung alle Dateien, wertet nur bestimmte Teile der HTML-Seiten aus und hängt alle verlinkten PDF- oder MS-Office-Dateien im gleichen Domino-Datensatz an, manchmal über 20 Dateien. Durch ein Kontrollverfahren können neue, geänderte oder gelöschte, nicht mehr vorhandene Seiten erkannt werden und der Index ist täglich aktuell. Auch die IBM Lotus Domino-Maschine läuft hier unter Linux. Parallel dazu läuft die maschinelle Indexierung als kontinuierlicher Prozess, denn diese Engine ist im Gesamtsystem der langsamste und rechenintensivste, weil jeder String gegen die linguistischen Wörterbücher geprüft werden muss und Autindex eine einfache Satzanalyse durchführt, die Phrasen erkennt und die Relevanz von Substantiven in Sätzen, Absätzen und Dokumenten in Abhängigkeit vom eigenen Fachthesaurus ermittelt.

5 eGovernment Klassifikation und Thesaurus

In der zweiten Phase wird zusätzlich der neue eGovernment-Thesaurus mit einer polyhierachischen Klassifikation noch umfassender zum Einsatz kommen als nur zur Query-Expansion. Der Thesaurus versucht, alle Lebensbereiche und alle Themenfelder terminologisch einzufangen und systemtatisch darzustellen. Die Vergabe der Klassen kann in Phase zwei entweder automatisiert vorgeschlagen werden, aufgrund der in der maschinellen Indexierung ermittelten wichtigen Begriffe oder durch manuelle Eingabe von Klassen oder Thesaurusbegriffen – es soll automatisch in beiden Richtungen dann ein Eintrag erfolgen.

Abbildung 7: Ein kleiner Ausschnitt aus dem Thesaurus in IC INDEX. Der Begriff ist mit einer Klasse, mehreren Synonymen, einem Ober- und mehreren Unterbegriffen vernetzt.

Das Ziel der klassifikatorischen Erschließung sind nach Klassen sortierte Suchergebnisse, so wie dies über Angebote von Amazon oder eBay vielen Web-Anwendern bekannt sind. Da die Menge der Dokumente im CMS noch überschaubar ist, wird es aber keine rein maschinelle Klassifizierung geben.

Technisch basiert das Thesaurusentwicklungsprogramm IC INDEX wieder auf IBM Lotus Notes & Domino. Die Integration von CMS-Daten und Thesaurus wird ganz in der Lotus Notes-Umgebung erfolgen.

5 Bewertung und Ausblick

Alle Anwendungen für bzw. mit dem Land Vorarlberg überwinden bei der Frageformulierung graduell Sprachgrenzen durch zusätzliche fachlich relevante Synonyme, Unterbegriffe und Übersetzungen. Sie wiederum treffen auf sprachlich normalisierte Dokumente, was in der deutschen Sprache nur mit leistungsstarken linguistischen Verfahren möglich ist. Das Relevance Ranking ist zusätzlich „getunt", d.h. Felder und bestimmte Feldwerte werden „vorgezogen". Dies trifft zu ausser bei LIS, dort greifen bisher nur die Standard-Verfahren der GTR, der SearchEngine in Lotus Domino, mittelfristig sollte auch hier stärker getunt werden. Die Nachfrage und Benutzerakzeptanz sprechen bei dandelon.com und LIS eindeutig für den gewählten Ansatz. Die Fortschritte bei www.vorarlberg.at sind, wer die vorhergende Lösung kennt, unübersehbar. Und: Für den Anwender – ist das alles „ganz normal".

Aktuell bereitet der Autor einen Technologie-Wechsel vor, indem die SearchEngine GTR in IBM Lotus Domino durch IBM Omnifind ersetzt wird. Dann werden linguistisch noch wesentlich mehr Sprachen unterstützt und weitere Datensammlungen können noch schneller erschlossen werden.

URL-Verzeichnis

dandelon.com
http://www.dandelon.com
Mit einigen kompletten Aufsätzen zu dandelon.com im Bereich „Artikel" und weitere Aufsätze in FAQ-Bereich

Landtagsinformationssystem Vorarlberg:
http://www.vorarlberg.at/landtag/landtag/recherche/recherche.htm

Gemeindedokumentation Österreich
http://www.gemeindedokumentation.at

Landeshomepage
http://www.vorarlberg.at